# Canonical Correlation Analysis

Uri Shaham

March 4, 2024

## 1 Rayleigh quotient

**Definition 1.1.** *Let $A \in \mathbb{R}^{m\times m}$ be a symmetric matrix and let $x \in \mathbb{R}^m$ be a nonzero vector. Their Rayleigh quotient is defined as*

$$R(A,x) = \frac{x^T Ax}{x^T x}. \tag{1}$$

**Theorem 1.2.** *Let $A = V\Lambda V^T$ be the eigendecomposition of $A$, where $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_m)$, and $\lambda_i \geq \lambda_j$ for $i < j$. Then $\max_x R(A,x) = \lambda_1$, the largest eigenvalue of $A$ and the corresponding eigenvector $v_1$ is the maximizer.*

*Proof.* Let $A = V\Lambda V^T$ be the eigendecomposition of $A$. Define $y = V^T x$. Then

$$\begin{aligned} R(A,x) &= \frac{x^T Ax}{x^T x} \\ &= \frac{x^T V\Lambda V^T x}{x^T VV^T x} \\ &= \frac{y^T \Lambda y}{y^T y}. \end{aligned} \tag{2}$$

Hence, to maximize the R(A, x) one needs to find a unit vector $y$ which maximizes $y^T\Lambda y$.

$$y^T\Lambda y = \sum_{i=1}^m y_i^2\lambda_i \leq \lambda_1 \sum_{i=1}^m y_i^2 = \lambda_1.$$

Observe that $y = (1,0,\ldots,0)^T$ gives $y^T\Lambda y = \lambda_1$. Hence $x = Vy = v_1$. □

## 2 Recap: principal component analysis

Let $X = (X_1,\ldots,X_m)^T \in \mathbb{R}^m$ be a random vector with zero mean and covariance $\Sigma$. In PCA we seek for a projection $w^T X$ of $X$ onto a direction $w$ along which the variance is maximized. This is the first principal direction. For each subsequent direction, we seek to maximize the variance and in the orthogonal complement of the subspace of previously selected components.

Recall that for a random vector $Z$ and a constant vector $a$, both in $\mathbb{R}^m$, $\mathrm{Var}(a^T x) = a^T\,\mathrm{Cov}(Z)a$. Thus, formally, to find the first principal direction $w$, we maximize

$$\mathrm{Var}\left(\frac{w^T}{\|w\|}X\right) = \frac{w^T\Sigma w}{w^T w}. \tag{3}$$

Equation (3) is a Raighley quotient and hence its maximizer is the vector $w \in \mathbb{R}^m$ which is the eigenvector of the covariance matrix $\Sigma$ with the largest eigenvalue (which we will denote by $v_1(\Sigma)$). In the homework, you will prove that since all principal components are orthogonal, subsequent directions are the next eigenvectors.

Let $X_n \in \mathbb{R}^{n\times m}$ be a collection of $n$ i.i.d samples of $X^T$, and let $\Sigma_n = \frac{1}{n}X_n^T X_n$ be the sample covariance matrix. Let $X_n = U\Lambda V^T$ be the SVD of $X_n$, so $\Sigma_n = V(\Lambda^2/n)V^T$ is the eigendecomposition of $\Sigma_n$. The projection of the data onto the principal directions is then $X_n^T V$. To map the data back to the original coordinates one simply multiplies the projected data matrix from the right by $V^T$. Observe the PCA projection is $X_n V = U\Lambda$, thus PCA and SVD (on centered data) are in fact equivalent procedures.

# 3 Aside: PCA signal recovery in spike models

Let $X$ be $m$-dimensional multivariate normal random vector, with zero mean. Assume the covariance $\Sigma = \mathbb{E}(XX^T) = I_d + \beta uu^T$, where $u \in \mathbb{R}^d$ is a unit vector. Thev $I_d$ component of the covariance can be viewed as noise (symmetric to all directions). We view $u$ as a signal, indicating a high variance direction in the data, and $\beta \geq 0$ is the signal strength Let $\Sigma_n$ be the sample covariance matrix. Observe that the largest eigenvalue os $\Sigma$ is $(1+\beta)$, and the corresponding eigenvector is $u$.

A result due to Baik, Ben-Arous and Peche (2005) specifies conditions on the recovery of the signal as the sample size $n$ approaches infinity. Specifically, let $r = \frac{m}{n}$. Then

$$\mathbb{P}\left\{\lim_{n\to\infty}\lambda_1(\Sigma_n) = \begin{cases}(1+\sqrt{r})^2, & \text{if } \beta \leq \sqrt{r}\\ (\beta+1)(1+\frac{r}{\beta}), & \text{if } \beta \geq \sqrt{r}\end{cases}\right\} = 1.$$

In addition,

$$\mathbb{P}\left\{\lim_{n\to\infty}\langle v_1(\Sigma_n), u\rangle = \begin{cases}0, & \text{if } \beta \leq \sqrt{r}\\ \frac{1-\frac{r}{\beta^2}}{1+\frac{r}{\beta^2}}, & \text{if } \beta \geq \sqrt{r}\end{cases}\right\} = 1.$$

Together, these results imply that PCA recovers the largest component iff $n \geq \frac{m}{\beta^2}$. Specifically, the sample size needs to be linear in the dimension, and as the signal is stronger, a smaller sample may suffice.

# 4 Canonical correlation analysis

Let $X \in \mathbb{R}^d$, $Y \in \mathbb{R}^m$ be random vectors. In CCA we seek for linear combinations of $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_m)$ which are maximally correlated. In the sample version, let $X_n \in \mathbb{R}^{n\times d}$ and $Y_n \in \mathbb{R}^{n\times m}$ be collections of $n$ samples from each random variable, such that for $i \neq j$, $(x_i, y_i)$ (i.e., the $i$'th sample from $(X, Y)$) and $(x_j, y_j)$ are independent. CCA seeks orthogonal bases for $\text{col}(X_n)$ and $\text{col}(Y_n)$ such that the cross-correlations are maximized.

Let's start with finding the first pair of projections, $U := a^T X, V := Y^t b$. Assume that $X$ and $Y$

both have zero mean. Then

$$\begin{aligned}\rho_1 &:= \operatorname{corr}(U, V) \\ &= \frac{\mathbb{E}[UV]}{\sqrt{\mathbb{E}[U^2]}\sqrt{\mathbb{E}[V^2]}} \\ &= \frac{\mathbb{E}\left[a^T XY^T b\right]}{\sqrt{\mathbb{E}\left[a^T XX^T a\right]}\sqrt{\mathbb{E}\left[b^T YY^T b\right]}} \\ &= \frac{a^T \Sigma_{XY} b}{\sqrt{a^T \Sigma_X a}\sqrt{b^T \Sigma_Y b}},\end{aligned}$$

where $\Sigma_X, \Sigma_Y, \Sigma_{XY}$ are the covariance and cross-covariance matrices.

Next, we change the basis and define $c := \Sigma_X^{\frac{1}{2}} a,\ d := \Sigma_Y^{\frac{1}{2}} b$. Then

$$\rho_1 = \frac{c^T \Sigma_X^{-\frac{1}{2}} \Sigma_{XY} \Sigma_Y^{-\frac{1}{2}} d}{\sqrt{c^T c}\sqrt{d^T d}}. \tag{4}$$

Applying Cauchy-Schwartz inequality ($x^T y \leq \|x\|\|y\|$) on the numerator of equation (4) we have

$$\left(c^T \Sigma_X^{-\frac{1}{2}} \Sigma_{XY} \Sigma_Y^{-\frac{1}{2}}\right) d \leq \left(c^T \Sigma_X^{-\frac{1}{2}} \Sigma_{XY} \Sigma_Y^{-\frac{1}{2}} \Sigma_Y^{-\frac{1}{2}} \Sigma_{YX} \Sigma_X^{-\frac{1}{2}} c\right)^{\frac{1}{2}} \sqrt{d^T d},$$

where equality holds iff $X_Y^{-\frac{1}{2}} \Sigma_{YX} \Sigma_X^{-\frac{1}{2}} c$ and $d$ are in the same direction. Hence in that case

$$\rho_1^2 = \frac{c^T \Sigma_X^{-\frac{1}{2}} \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX} \Sigma_X^{-\frac{1}{2}} c}{c^T c}. \tag{5}$$

Equation (5) is Rayleigh quotient, hence its maximizer is $c = v_1(\Sigma_X^{-\frac{1}{2}} \Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX} \Sigma_X^{-\frac{1}{2}})$ (i.e., the eigenvector corresponding to the largest eigenvalue). $d$ is then obtained as a unit vector in the direction of $X_Y^{-\frac{1}{2}} \Sigma_{YX} \Sigma_X^{-\frac{1}{2}} c$. Similarly, by reversing the order of x and y in the above process we get that $d = v_1(\Sigma_Y^{-\frac{1}{2}} \Sigma_{YX} \Sigma_X^{-1} \Sigma_{XY} \Sigma_Y^{-\frac{1}{2}})$, and $c$ is then obtained as a unit vector in the direction of $X_X^{-\frac{1}{2}} \Sigma_{XY} \Sigma_Y^{-\frac{1}{2}} d$. Finally, reversing the change of variables we have $a = \Sigma_X^{-\frac{1}{2}} c$ and $b = \Sigma_Y^{-\frac{1}{2}} d$. The projected variables are then $U = a^T X$ and $V = b^T Y$. To obtain the next pairs $U_i, V_i, i = 2, \ldots, \min\{m, n\}$, we would like each new canonical directions to be uncorrelated with previous ones, hence the subsequent eigenvectors are used (see homework).